1988

The use of observed data for the initial-value problem in numerical weather prediction

Franke, R.

Comput. Math. Applic. Vol. 16, No. 1/2, pp. 169–184, 1988
Printed in Great Britain

0097-4943/88 $3.00 + 0.00
Pergamon Press plc

THE USE OF OBSERVED DATA FOR THE INITIAL-VALUE PROBLEM IN NUMERICAL WEATHER PREDICTION

R. FRANKE,[1] E. BARKER[2] and J. GOERSS[2]

[1]Department of Mathematics, Naval Postgraduate School and [2]Naval Environmental Prediction Research Facility, Monterey, CA 93943, U.S.A.

Abstract—The problem of combining observed and predicted values of meteorological variables, all with error, to obtain current weather conditions is considered. Statistical interpolation is in common use for this problem. Properties of isotropic spatial covariance functions are developed. The performance of several families of covariance functions in fitting published data is investigated. The second-order autoregressive covariance function is identified as having suitable theoretical and excellent approximation properties. Sensitivity of the errors in statistical interpolation to misspecification of the statistical parameters is explored, showing that the process is quite stable under such perturbations.

1. INTRODUCTION

The physical processes of the earth's atmosphere can be modeled by a system of hydrodynamic equations. This system of equations cannot be solved directly unless many simplifying assumptions are made, severely limiting how realistically the actual atmospheric processes are simulated. In order to produce accurate weather forecasts the full system of equations must be solved in four dimensions. In practice, global weather forecasts are produced at various meteorological centers around the world by treating these equations as an initial-value problem and integrating forward in time to produce forecasts. The solution of this problem requires the use of advanced vector processors as the number of computations involved is staggering. The forecasting problem was formulated quite succinctly by the Norwegian pioneer in weather forecasting, V. Bjerknes, when he defined the necessary and sufficient conditions for a successful system in an article written in 1922 [1] to be:

(i) A sufficiently accurate knowledge of the state of the atmosphere at the initial time.
(ii) A sufficiently accurate knowledge of the laws according to which one state of the atmosphere develops from another.

Bjerknes' discoveries of the hydrodynamical nature of the weather problem led several European nations, especially the Scandinavian countries, to begin collecting observations of the state of the atmosphere. This data collection led L. F. Richardson [2] to try describing initial conditions from a hand analysis and projecting the state of the atmosphere to the future from the hydrodynamical equations. The task was monumental as Richardson estimated that a warehouse of 64,000 people using the mechanical calculator of the day could just forecast the state of the atmosphere at the rate that it was actually evolving. Unfortunately, many factors, discovered later in the 1940s, kept Richardson from making a successful forecast.

The magnitude of the weather forecasting problem required the development of electronic computers for even simple solutions. The Electronic Numerical Integrator and Computer (ENIAC) developed in the late 1940s allowed Charney *et al.* [3] to succeed in making a reasonable 24 h forecast. Their hydrodynamical model was simplified to filter gravity waves while allowing weather patterns to develop in a manner similar to that observed in the atmosphere. Their initial conditions of pressure heights were derived by hand and the result gridded and typed into the computer.

With the rapid development of computers over the past 30 years, it has become possible to use numerical techniques to integrate the full set of hydrodynamic equations forward in time to

produce improved weather forecasts. As Bjerknes predicted, accurate forecasts require more than just accurate treatment of the physical processes of the atmosphere, they also require accurate specifications of the initial state from nonuniformly located observations. Panofsky [4], Bergthorsson and Döös [5] and Cressman [6] pioneered methods to use the computer to obtain a weather analysis from observational data. This process of combining observation values with a background field is called objective analysis. For the most part, these original objective analysis techniques were weighted average schemes that depended upon proper specification of several parameters, usually obtained in an *ad hoc* way. Today, most of the world's weather centers use statistical objective analysis techniques based on the work of Gandin [7] to provide initial conditions for their atmospheric forecast models.

In practice two sources of information are combined to produce an objective analysis: observations of atmospheric variables and a forecast made by the atmospheric model from a previous analysis. The forecast is commonly called the "first guess" to the analysis or the "background". Because the forecast is hardly a guess, the term "forecast background" is used in this paper to emphasize that the background to the analysis was derived from a forecast made earlier. The observations of temperature, wind and moisture are made by *in situ* instruments attached to balloons, aircraft and ships, or from remote instruments aboard satellites or on the earth's surface. The result is a collection of observations of varying degrees of accuracy taken at various times. The statistical analysis schemes have been designed to "optimally" combine these observations with the forecast background to produce the initial conditions required by the numerical forecast model. The optimality of these schemes directly depends upon how well the statistical properties of the errors of the forecast background and the observations are defined. In practice, the schemes are multivariate in the sense that they are used to simultaneously analyze multiple related dependent variables from measured values.

In this paper we will deal with the representation of the statistics of the forecast background error. In particular, the modeling of the spatial autocovariance of the error for the primary variable is examined. Early versions of Gandin's method used a simple exponential function to model the autocovariance. Although this model is simple, it failed to be sufficiently flexible to describe details of the statistics derived from actual data. A search of the literature revealed that many models are available, but tests of their abilities in fitting background statistics for an actual forecast model have not been conducted. The mathematical and precision limitations of various models have been determined and are described in this paper.

Optimum interpolation (OI), which is sometimes more practically referred to as statistical interpolation (SI), is applied to compute the corrections to the background field. This is done by first interpolating the background to the nonuniformly located observation locations, and then computing the difference between the observed and background value. If observations were exact, this would be the background error measured at discrete locations. These measurements of background error are then used by OI to compute a correction field on a uniform grid, which is added to the background to produce the analysis.

The full development of the equations for a multivariate application of OI is given in several papers [e.g. 8–14]. A brief outline of the method is given in the following. For a collection of estimates of the error at scattered points, it is desired to estimate the value of the error at the grid points. OI approximates these values by a linear combination of the known values defined so that the expected mean squared error over some ensemble of realization is minimized. This requires that the statistical properties (covariances between variables) be known. Stationarity (independence of the particular grid point) of the statistical parameters is required for a tractable problem. The weights used in the linear combination are obtained from the solution of a certain system of linear equations, the coefficient matrix being the matrix of covariances between the background plus observed errors at the observation points. The positive definiteness of this matrix plays an important role, both theoretically and computationally.

A discussion of multivariate covariance functions, properties they must satisfy and methods of obtaining such functions are discussed in Section 2. Experiments with several families of covariance functions in fitting background error statistics and the resulting performance in a statistical interpolation scheme are described in Section 3. Section 4 summarizes the results and suggests some future work.

2. MULTIVARIATE COVARIANCE FUNCTIONS

2.1. General development

The theory of covariance functions and that of positive definite functions go hand-in-hand. Positive definiteness of matrices such as occur in our application are equivalent to the spatial covariance function for the background errors being positive definite. Positive definite functions are characterized by Bochner's theorem [15], which states that a function is positive (semi)definite if and only if its Fourier transform is nonnegative. Alternatively, the covariance function is the Fourier (cosine) transform of a probability density (nonnegative) function. Because of the application, our interest is in positive definite functions that are smooth in the sense that certain partial derivatives exist. An excellent reference for positive definite functions is Stewart [16].

For completeness, a derivation of the covariance functions for variables related through differentiation is given here. Suppose that it is wished to analyze three related dependent variables, requiring that the corrections obtained via OI (or more correctly, SI) will not upset the relationship between the predicted values of the variables. Let the error in the predicted variables be denoted by $Z(x, y)$, $X(x, y)$ and $Y(x, y)$, where (x, y) gives the spatial location and it is assumed that $X(x, y) = k_1 Z_x(x, y)$ and $Y(x, y) = k_2 Z_y(x, y)$. The subscripts x and y denote partial differentiation with respect to x and y, respectively. Assume that the errors in the predicted values are stationary, i.e. the statistics do not depend on (x, y), and have zero mean. Using $\mathrm{E}[\cdot]$ to denote the expected value, or ensemble average, the spatial covariance function for Z, as a function of "lags" s and t, is

$$R(s, t) = \mathrm{E}[Z(x, y)Z(x + s, y + t)] = \mathrm{E}[Z(x - s, y - t)Z(x, y)].$$

The latter equality follows from stationarity. Under the assumption that the order of partial differentiation and the expected value can be interchanged, the cross covariance functions and the covariance functions for the derived variables are found in the manner illustrated here (of course, it is assumed throughout this paper that the necessary derivatives exist):

$$\begin{aligned}\mathrm{E}[Z(x, y)X(x + s, y + t)] &= \mathrm{E}[Z(x, y)k_1 Z_x(x + s, y + t)] \\ &= \mathrm{E}[Z(x, y)k_1 Z_s(x + s, y + t)] = k_1 \mathrm{E}[Z(x, y)Z(x + s, y + t)]_s \\ &= k_1 R_s(s, t),\end{aligned}$$

while

$$\begin{aligned}\mathrm{E}[X(x, y)Y(x + s, y + t)] &= \mathrm{E}[X(x, y)k_2 Z_y(x + s, y + t)] \\ &= \mathrm{E}[X(x, y)k_2 Z_t(x + s, y + t)] = k_2 \mathrm{E}[X(x, y)Z(x + s, y + t)]_t \\ &= k_2 \mathrm{E}[k_1 Z_x(x - s, y - t)Z(x, y)]_t = k_2 \mathrm{E}[-k_1 Z_s(x - s, y - t)Z(x, y)]_t \\ &= -k_1 k_2 \mathrm{E}[Z(x, y)Z(x + s, y + t)]_{ts} = -k_1 k_2 R_{ts}(s, t).\end{aligned}$$

Note that while the covariance functions are symmetric, the cross covariance functions are antisymmetric, which accounts for the sign change that comes from changing the order of the product in the expected value. This means, among other things, that the cross covariance must be zero at zero lag values. This behavior can be seen in the function plots of Bergman [11] and Schlatter *et al.* [10].

2.2. Some necessary properties

In order for the covariances of the derived functions and the cross covariance functions to exist, certain conditions must be satisfied by the function $R(s, t)$. These have been alluded to by Buell [17], and are given by Julian and Thiébaux [18], where

$$\lim_{s \to 0} \frac{R_s(s)}{s} \quad \text{is finite, and} \quad \lim_{s \to 0} \left[\frac{R_s(s)}{s} - R_{ss}(s) \right] = 0.$$

In this equation s represents the lag distance [lag in the above was $(s^2 + t^2)^{1/2}$], and $R(s)$ is an isotropic covariance function. When one considers that $R_s(0)$ must be zero, the first limit is the

definition of the second derivative at $s = 0$, hence existence of the limit means that the covariance function must be twice differentiable at $s = 0$. The second limit then says that the second derivative is continuous at $s = 0$. Thus the theorem given by Julian and Thiébaux [18] can be simplified:

> **Theorem 1.** If $R(s)$ is an isotropic covariance function for Z in two dimensions, then the covariance functions for the partial derivatives of Z exist at $s = 0$ iff $R(s)$ is twice continuously differentiable at $s = 0$.

2.3. Anisotropic functions

It has been contended that isotropic covariance functions do not adequately model the forecast error statistics and that gains can be made by using anisotropic functions. (See works by Thiébaux and coworkers [13, 19–21] for development and discussion of product forms of covariance functions.) Use of products of single-dimensional functions has the advantage of carrying over desirable properties to higher dimensions, as well as being able to use essentially one-dimensional structures and techniques. On the other hand, perusal of contour plots of product functions show that zero crossings of the functions occur along grid lines, and it is easy to see this will always happen. This may be undesirable behavior, and almost certainly it is not the kind of behavior seen in the error statistics.

Another form of anisotropy is possible, one which results from scaling differently in two orthogonal directions, then using an isotropic function in the scaled variables. This would result in the zero crossings in the contour plots of the function being ellipses with axes in the two directions, and all contours having the same shape. The eccentricity of the ellipse is a measure of the anisotropy of the error statistics. It would be easy to allow rotation along with the scaling to obtain ellipses of constant "distance" with any axis orientation. For a discussion of this type of anisotropic correlations, see Seaman [22] and Buell and Seaman [23]. The properties of any such functions are those of isotropic functions, of course, since the anisotropy arises purely from a rotation and scaling.

2.4. Isotropic functions

The use of isotropic functions in two or more dimensions that have been derived from one-dimensional considerations can possibly lead to nonpositive definite functions. For example, Ripley [24, p. 11] quotes a result of Matèrn [25], which gives a lower bound for isotropic positive definite functions in several dimensions. The result means that positive definite functions in two dimensions are necessarily bounded below by -0.403, the minimum value of $J_0(s)$, while in three dimensions the bound is -0.218. Thus any oscillatory positive definite function in one dimension that takes on values < -0.403 cannot be an isotropic positive definite function in two dimensions. A positive definite function with parameters to separately control the oscillation frequency and the decay can probably be made into a nonpositive definite isotropic function in two dimensions. For example, an exponentially damped cosine function, $f(s) = \cos(as)\exp(-bs)$, can be made nonpositive definite by suitable choice of parameters, say $a = 5$ and $b = 0.1$. This result also applies to other candidates for isotropic correlation function models, as will be shown later.

There is a one-to-one correspondence between covariance functions in one dimension and isotropic covariance functions in two dimensions. Using the so-called "turning band" method, Matheron [26] presents a way of generating an isotropic d-dimensional covariance function from a one-dimensional covariance function. The relation is

$$C_d(s) = \mathrm{K} \int_0^1 C_1(vs)(1 - v^2)^{(d-3)/2}\, dv,$$

where K is a constant that is unimportant for our purposes. In two dimensions, this gives

$$C_2(s) = \mathrm{K} \int_0^1 C_1(vs)(1 - v^2)^{-1/2}\, dv.$$

It is possible to invert Matheron's relation to show a one-to-one relationship. A sketch of the inversion process follows. Employing a change of variables in the previous expression gives

$$C_2(s) = \mathrm{K} \int_0^s C_1(t)(s^2 - t^2)^{-1/2}\, dt,$$

then making further change of variables, $s^2 = x$ and $t^2 = y$, yields

$$C_2(x^{1/2}) = \mathrm{K} \int_0^x [C_1(y^{1/2})(2y)^{-1/2}](y - x)^{-1/2}\,\mathrm{d}t.$$

This is Abel's equation for $\mathrm{K}C_1(y^{1/2})(2y)^{-1/2}$, and the well-known solution [27] is given by

$$C_1(x^{1/2}) = \mathrm{K}'x^{1/2}\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x C_2(y^{1/2})(x - y)^{-1/2}\,\mathrm{d}y,$$

where K′ is a different constant. Substituting for s and t once again, gives

$$C_1(s) = \mathrm{K}'s\frac{\mathrm{d}}{\mathrm{d}(s^2)}\int_0^s C_2(t)(s^2 - t^2)^{-1/2}2t\,\mathrm{d}t.$$

The correspondence between covariances in one dimension and three dimensions is easier to invert, and is given by Ripley [24]. There the relation is

$$C_3(s) = \int_0^1 C_1(vs)\,\mathrm{d}v \quad \text{and} \quad C_1(s) = \frac{\mathrm{d}}{\mathrm{d}s}[sC_3(s)].$$

While this characterization of multidimensional isotropic covariance functions is interesting, and can in fact be used to generate isotropic multidimensional covariance functions, it does not easily answer the question as to whether or not a particular one-dimensional function is an isotropic positive definite function in more dimensions. One way to answer such a question is to use the characterization of positive definite functions as Fourier transforms of probability density functions (or alternatively, as functions whose Fourier transform is positive). The Fourier transform of an isotropic function $C(s)$ in two dimensions becomes (essentially) the Hankel transform of $s^{1/2}C(s)$. It may be considerably easier to look at the one-dimensional Fourier transform. It would then be useful to have a sufficient condition on the Fourier transform of the function which would guarantee it is an isotropic positive definite function in two dimensions. Such a condition will now be derived. Let $C_1(s)$ be a positive definite function of one variable. From the characterization in the previous section, it can be shown that

$$C_1(s) = \int_0^\infty \cos(rs)h(r)\,\mathrm{d}r, \tag{1}$$

for some probability density function $h(r)$ (i.e. $h(r) \geqslant 0$, with integral equal to one). The problem is then to determine the conditions that will make $C_1(s)$ the two-dimensional Fourier transform of an isotropic probability density function. Such a transform is necessarily isotropic. A function $g(s)$ is sought so that

$$C_1(r) = \int_0^\infty J_0(rs)sg(s)\,\mathrm{d}s. \tag{2}$$

This expression is inverted using the Hankel transform, giving

$$g(s) = \int_0^\infty J_0(sr)rC_1(r)\,\mathrm{d}r. \tag{3}$$

Then, using equation (1) in equation (3), and interchanging the order of integration, followed by integration by parts yields

$$\begin{aligned} g(s) &= \int_0^\infty J_0(sr)r\left(\int_0^\infty \cos(tr)h(t)\,\mathrm{d}t\right)\mathrm{d}r \\ &= \int_0^\infty h(t)\left(\int_0^\infty J_0(sr)r\cos(tr)\,\mathrm{d}r\right)\mathrm{d}t \\ &= \int_0^\infty h(t)\left(-\frac{\mathrm{d}}{\mathrm{d}t}\int_0^\infty J_0(sr)\sin(tr)\,\mathrm{d}r\right)\mathrm{d}t \\ &= \int_0^\infty h'(t)\left(\int_0^\infty J_0(sr)\sin(tr)\,\mathrm{d}r\right)\mathrm{d}t, \end{aligned}$$

and then

$$g(s) = -\int_0^\infty h'(t)/(t^2 - s^2)^{1/2}\,dt. \tag{4}$$

The last equality uses the Hankel transform of $r^{-1/2}\sin(tr)$.

In order for $g(s)$ to be a probability density function it must be nonnegative with integral equal to one. It is easy to show (again, interchanging the order of integration) that the integral is equal to one. It is more difficult to show necessary and sufficient conditions for the nonnegativity of $g(s)$. The above relations summarized give:

Theorem 2. A sufficient condition for $C_1(s)$ to be a valid isotropic covariance function in two dimensions is that $h(t)$ be a monotone decreasing ($h'(t) \leqslant 0$) function.

This condition seems unnecessarily restrictive, and difficult to use since the condition is on the Fourier cosine transform of $C_1(s)$ rather than $C_1(s)$ itself. Nonetheless, the condition can be used to show the following interesting results.

I. Consider the exponentially damped cosine function,

$$C(s) = \cos(as)e^{-bs}.$$

The Fourier cosine transform of this function is

$$h(t) = \mathcal{F}(C)(t) = \frac{b(b^2 + a^2 + t^2)}{[b^2 + (a-t)^2][b^2 + (a+t)^2]}.$$

Inspection of $h'(t)$ shows that if $b^2 \geqslant 3a^2$, it is nonpositive $\forall t$, and hence $h(t)$ is monotone decreasing under that constraint.

II. Consider the second-order autoregressive (SOAR) covariance function,

$$C(s) = [\cos(as) + (b/a)\sin(as)]e^{-bs}.$$

The Fourier cosine transform of this function is

$$h(t) = \mathcal{F}(C)(t) = \frac{2b(b^2 + a^2)}{[b^2 + (t-a)^2][b^2 + (t+a)^2]}.$$

Inspection of $h'(t)$ reveals that if $b^2 \geqslant a^2$, it is nonpositive $\forall t$, and thus $h(t)$ is monotone decreasing under that constraint. We see that each of the above $C(s)$ is an isotropic positive definite function, hence is a covariance function if the appropriate inequality on the parameters is satisfied.

III. Consider the special case of the damped cosine function,

$$C(s) = [A + (1-A)\cos(as)]/[1 + (bs)^2]^{1/2}.$$

The Fourier transform of this function is

$$h(t) = \mathcal{F}(C)(t) = (2b)^{-1}\{(1-A)[K_0(|t-a|/b) + K_0(|t+a|/b)] + AK_0(t/b)\}.$$

Because the modified Bessel function K_0 becomes unbounded as the argument tends to zero, for $A \neq 1$, the Fourier transform must be increasing as $t \to a$ through values $< a$. For $t > a$, and possibly for some values $< a$, the function is decreasing. Thus the sufficient conditions given above are not met, and it is easy to find configurations of (x, y) points and parameter values A, a and b for which the resulting "covariance" matrix is not positive definite. The two-dimensional Fourier transform of $C(s)$ [the Hankel transform of $s^{1/2}C(s)$] has thus far gone unsolved, so it is presently unknown if there are parameter values (other than for $A = 1$) that will yield a positive definite function.

IV. Consider the Bessel function $J_0(as)$. The Fourier transform of this function is

$$h(t) = \mathcal{F}(C)(t) = \begin{cases} 0, & t < a \\ t^{1/2}/(t^2 - a^2)^{1/2}, & t > a \end{cases}.$$

This function is easily seen to be monotone decreasing for $t > a$, and thus the Bessel function $J_0(as)$ is an isotropic covariance function in two dimensions. Application of this relation requires attention to some technical details because of the infinite jump discontinuity at $t = a$.

The above results concerning several functions proposed for use as isotropic covariance functions in two dimensions are useful. The lack of results and empirical evidence against the damped cosine being positive definite negate the results noted in the next section where we see that the fitting power of the function is very good. These aspects of the function will be discussed further in the next section.

2.5. Summary

This section contains some useful information for the construction of isotropic positive definite functions and testing of functions for positive definiteness. When possible, the two-dimensional Fourier transform of $C_1(s)$ can be used to decide whether or not the function is positive definite. When the two-dimensional Fourier transform cannot be obtained in closed form, Theorem 2 can give some information if the one-dimensional Fourier transform is available in closed form. While the sufficient condition given by Theorem 2 is not necessary, it has been shown to be useful in investigating some functions which have been proposed for use as isotropic covariance functions in two dimensions.

3. SOME EXPERIMENTS WITH ISOTROPIC COVARIANCE FUNCTIONS

3.1. Background

The work reported in this section is intended to help determine something about the overall fitting properties of various suggested covariance functions. The term "overall fitting properties" is meant to include not only the ability of the function to model a reasonably complicated true covariance function, but also its performance when used in a statistical interpolation scheme with several different observation patterns.

The approach for this project was to begin with published data from an actual case, and then construct a covariance function using a least-squares fit to the data from a certain class of covariance functions. This model is used to define the "truth" model. Functions from other classes were then fitted to the same data, again in the least-squares sense, and the performance of these "assumed" covariance functions measured against that of the optimum model. The results to be discussed give some insight into what classes of functions have adequate fitting ability for modeling actual forecast error statistics, and also show how much skill is lost (in the idealized case) by use of inaccurate covariance functions.

The results given here consist of representative plots of assumed correlation functions together with the correlation function defined as "truth", and contour plots of some of the resulting expected errors. The tables show expected root-mean-square (e.r.m.s.) errors (relative to the standard deviation of the background error) over three grids of points and associated observation locations. The expected errors were computed as in Ref. [22]. The results obtained with various assumed correlation functions in the SI scheme are discussed in detail.

Additional plots are given by Franke [28].

3.2. The model correlation function

The data for the covariance function was obtained (by hand) from Lonnberg [29]. The data taken was plotted points from a covariance function of the type used by the European Center for Medium-range Weather Forecasts, in this case a five-term (i.e. $n = 5$) Bessel series of the form

$$\sum_{i=1}^{n} A_i J_0(s^* k_i/R) + A_0, \tag{5}$$

where k_i is the ith zero of the Bessel function $J_0(s)$ and R is the radius of the region of interest. This function is positive definite as an isotropic function in two dimensions, provided the coefficients A_i are all positive. In the work of Lonnberg [29], $R = 2000$ km. In this work, distance was measured in degrees, and the radius was scaled to 30°.

Least-squares fit to the data by functions of the type (5) for four, five and six terms were computed. While the original paper [29] indicated that a series with five terms generated the data, it was found that six terms yielded all positive coefficients and a significant reduction in the residual over five terms. Thus, it was decided to adopt the six-term series as the "truth" covariance function.

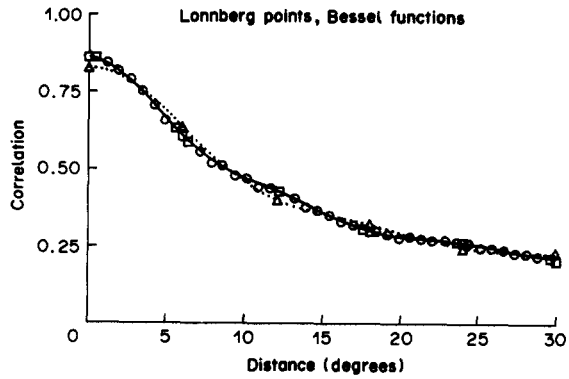

Fig. 1. The data points and least-squares fits by four- and six-term Bessel functions. (□) Six-term Bessel; (○) Lonnberg points; (△) four-term Bessel.

This six-term series would also be marginally harder to approximate using other classes of covariance functions. The data and the fits using four and six terms are shown in Fig. 1, and the coefficients are given in Table 1, along with other data. The intercept values of the approximations were 0.8270 and 0.8592 for four and six terms, respectively. This occurs because the data represents the spatial correlation of the background plus observation error, thus the intercept is a function

Table 1. Parameter values and e.r.m.s. errors; the model correlation is six-term Bessel

Assumed correlation function	Parameters		e.r.m.s.		
	a, b, c	A, A_i	MUS	EC	MA
Six-term Bessel		0.2474	0.2667	0.3752	0.7483
		0.3335			
		0.1844			
		0.1031			
		0.0362			
		0.0554			
		0.0400			
Four-term Bessel		0.2811	0.3046	0.4088	0.7503
		0.3090			
		0.2213			
		0.0930			
		0.0956			
NSE	10.0	0.0	0.3047	0.4184	0.7822
NSE	14.88	0.3200	0.3688	0.5282	0.7631
NSE	10.0	0.2500	0.3098	0.4158	0.7541
SOAR	0.0	0.0	0.3034	0.4022	0.7562
	0.1215				
SOAR	0.0	0.0	0.2931	0.3968	0.7562
	0.1374				
SOAR	0.0	0.2722	0.2780	0.3859	0.7491
	0.2055				
TOAR	0.4732	0.1974	0.2717	0.3794	0.7485
	0.3828				
	0.0914				
Damped cosine	0.4749	0.9592	0.2686	0.3779	0.7486
	0.1367				
	0.5000				
NSE[a]	15.0	0.0	0.3619	0.4414	0.7649
NSE[a]	12.31	0.3205	0.3474	0.4299	0.7593
SOAR[a]	0.0	0.3758	0.2743	0.3825	0.7495
	0.2654				
TOAR[a]	0.4468	−5.9965	0.2697	0.3801	0.7514
	0.1482				
	0.0052				
Damped cosine[a]	1.2236	1.0027	0.2749	0.3734	0.7491
	0.1507				
	0.5000				
Damped cosine	0.7009	1.0105	0.2692	0.3779	0.7484
	0.2069				
	0.3753				
Damped cosine[a]	0.7987	1.0147	0.2706	0.3784	0.7485
	0.2350				
	0.3317				

[a]These correlation functions were obtained by a least-squares fit over the interval (0°, 15°).

of the ratio of the standard deviations of background and observation error. The effects of this kind of discrepancy will be discussed in Section 3.3. The correlation function for background error is the approximation normalized to have value one at $s = 0$, of course.

3.3. The grid and observation point sets

Three grids and associated point sets were selected for studying the expected errors of statistical interpolation schemes based on various assumed covariance functions. All were based on the approximate locations of radiosonde data (from Refs [30, 31]) within the selected grid. Each grid covered a region that was 30° in longitude and 20° in latitude, and the three were chosen to represent a dense observation set, a partially dense observation set and a sparse observation set. The regions correspond to: the middle United States with 36 observations; the eastern United States and western Atlantic Ocean with 25 observations; and the middle Atlantic Ocean with 3 observations. For reference purposes, the three regions will be referred to as the MUS (mid-U.S.), EC (East Coast), and MA (mid-Atlantic) regions. The regions and the observation locations can be seen in Figs 2 and 3, parts (b), (c) and (d), respectively. The regions were gridded at 2.5° intervals for purposes of computing expected errors, although the e.r.m.s. errors given in Table 1 are only over the interior grid points to minimize edge effects. Use of interior grid points for this purpose is valid since on a sphere it is not necessary to interpolate to the boundary points. For contouring purposes the fields were interpolated to finer grids using bicubic spline interpolation.

3.4. The assumed correlation functions and results

The families of assumed correlation functions fell into five classes: (i) Bessel function; (ii) negative squared exponential (sometimes called Gaussian); (iii) autoregressive, of second order; (iv) autoregressive, third order; and (v) damped cosine. They will be discussed in turn, along with the results. Plots of the assumed correlation functions, along with the "truth" correlation function, are shown in Figs 2(a) and 3(a). For fitting purposes, each included a multiplicative parameter that determined the $s = 0$ intercept, and was subsequently dropped to obtain the correlation function. The value of this parameter is of interest, however, because dropping it shifts the curve (upward) to pass through the point (0, 1), and thus different fits may be shifted by differing amounts, which ultimately affects the fit to the background error correlation function.

Recall that the e.r.m.s. errors given in Table 1 are given as a fraction of standard deviation of the background error. The ratio of the standard deviation of the observation errors to the standard deviation of the background errors was 1/3.

(*i*) *Bessel function.* The reference expected errors were computed using the actual correlation function model, given by equation (5), with coefficients as given in Table 1. The results are given in Table 1, and are the smallest expected errors that can be obtained using a correction-to-background scheme, i.e. they are truly optimum. The correlation function is shown in Fig. 2(a), while the contour plots of the expected error for each of the three grid/observation sets is shown in Figs 2(b–d).

The results using a four-term Bessel function are given in Table 1. Because the intercept of the fit to the data is 0.8270 vs 0.8592, normalization to value one produces a curve that is predominantly above that for the model correlation function, especially for small distances. The result of the poor approximation for small distances is most pronounced over MUS. The effect was small over the sparse part of EC and over MA.

(*ii*) *Negative squared exponential (NSE).* The NSE has been recognized as inadequate for modeling error covariances for some time, and the results obtained here confirm that. The assumed form of the function was

$$A + (1 - A)\exp[-(s/b)^2]. \tag{6}$$

This function is positive definite as an isotropic function in two dimensions for $0 \leqslant A < 1$.

The initial fit was not obtained by least squares, but simply by attempting to fit the model correlation reasonably well for small distances, taking $A = 0$. The fit is reasonably good up to about 6°, and quite poor at greater distances. The expected errors are similar in magnitude to the expected errors for the four-term Bessel function, except over MA, where the errors are larger. However,

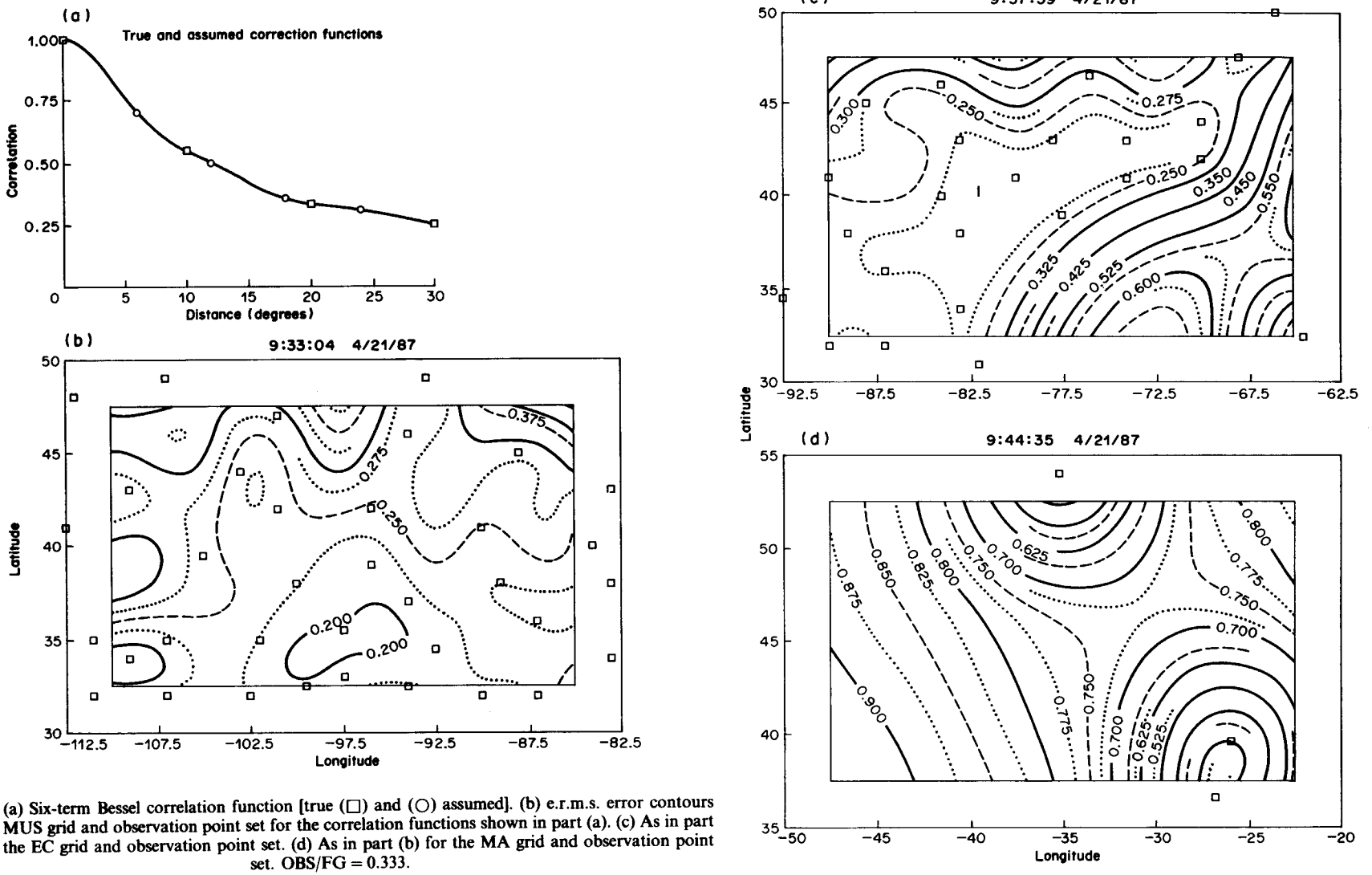

Fig. 2. (a) Six-term Bessel correlation function [true (□) and (○) assumed]. (b) e.r.m.s. error contours for the MUS grid and observation point set for the correlation functions shown in part (a). (c) As in part (b) for the EC grid and observation point set. (d) As in part (b) for the MA grid and observation point set. OBS/FG = 0.333.

since the errors over MA tend to be large anyway, the relative effect is not as great as one might expect.

The second attempt was by least squares for the parameters A and b. Because the NSE is too flat near the origin, this process yielded an intercept value of 0.8060, shifting the correlation function so that it is entirely above the model correlation curve. This results in even poorer performance over MUS and EC than the previous model, due to the inaccurate representation for small distances. The performance over MA was better than the above.

Due to the poor performance (compared to the above) obtained by adding a constant to the basic NSE it was decided to attempt to find a better fit by trial and error. No claim is made about any optimality for this function. The results in Table 1 demonstrate that it is probably not possible to obtain good results overall with a function from the NSE family, and certainly not for the present model correlation function.

(*iii*) *Autoregressive, second order* (*SOAR*). The SOAR model has been suggested as appropriate by Yudin [32] and Thiébaux [13] and this is supported by simulations due to Balgovind *et al.* [33]. This is the model that is being incorporated into the U.S. Navy NWP models. The formula given here includes a constant term which is not part of the SOAR model, but which has been noted to improve performance considerably [21]; and those results are confirmed here. The SOAR function with additive constant is

$$A + (1 - A)[\cos(as) + (b/a)\sin(as)]e^{-bs}. \tag{7}$$

This function is positive definite (in two dimensions) whenever $a \leqslant b$, and $0 \leqslant A \leqslant 1$. In all cases investigated here, and as has been reported elsewhere [e.g. 21], the parameter a tends to be essentially zero. In this case the function reduces to

$$A + (1 - A)[1 + bs]e^{-bs}. \tag{7a}$$

The initial attempt was a least-squares fit to the data with $A = 0$. The intercept obtained was 0.7977, with the resulting correlation curve then being considerably above the model correlation curve between 0° and 15°. The performance was only slightly better than with any of the previous correlation functions. It was then decided to attempt a least-squares fit with the intercept constrained to be 0.8592, the same as obtained for the model correlation function, but again with $A = 0$. Table 1 shows marginal improvement for all three grid/observation patterns. A third attempt included A in the least-squares fit, with no constraint. This resulted in a much closer match to the model correlation function, although the intercept of 0.8441 moved the assumed correlation curve above the model curve for much of the interval. The fit and resulting expected error contours are shown in Figs 3(a–d). Table 1 shows there is considerable improvement over all previous results, the most improvement being for MUS, and the least for MA.

(*iv*) *Autoregressive, third order* (*TOAR*). The use of the TOAR model has been investigated by Thiébaux *et al.* [21], including an additive constant. The formula is

$$A + (1 - A)\{[\bar{a}\cos(as) + \bar{b}\sin(as)]e^{-bs} + \bar{c}e^{-cs}\}, \tag{8}$$

where the coefficients $\bar{a}$, $\bar{b}$ and $\bar{c}$ are functions of a, b and c, given by

$$\bar{a} = (3b^2 - a^2 - c^2)ac/D,$$

$$\bar{b} = (b^2 - 3a^2 - c^2)bc/D$$

and

$$\bar{c} = -2(b^2 + a^2)ab/D,$$

where

$$D = (3b^2 - a^2 - c^2)ac - 2(b^2 + a^2)ab.$$

It is unknown what restrictions (beyond $0 \leqslant A \leqslant 1$) on the parameters are required to ensure the function is positive definite as an isotropic function in two dimensions.

The data was fitted by least squares with the TOAR function (8). The intercept was 0.8651, which resulted in the curve being slightly below the model correlation curve over most of the range. Overall, the fit was quite close and better than any of the previously discussed functions. The results

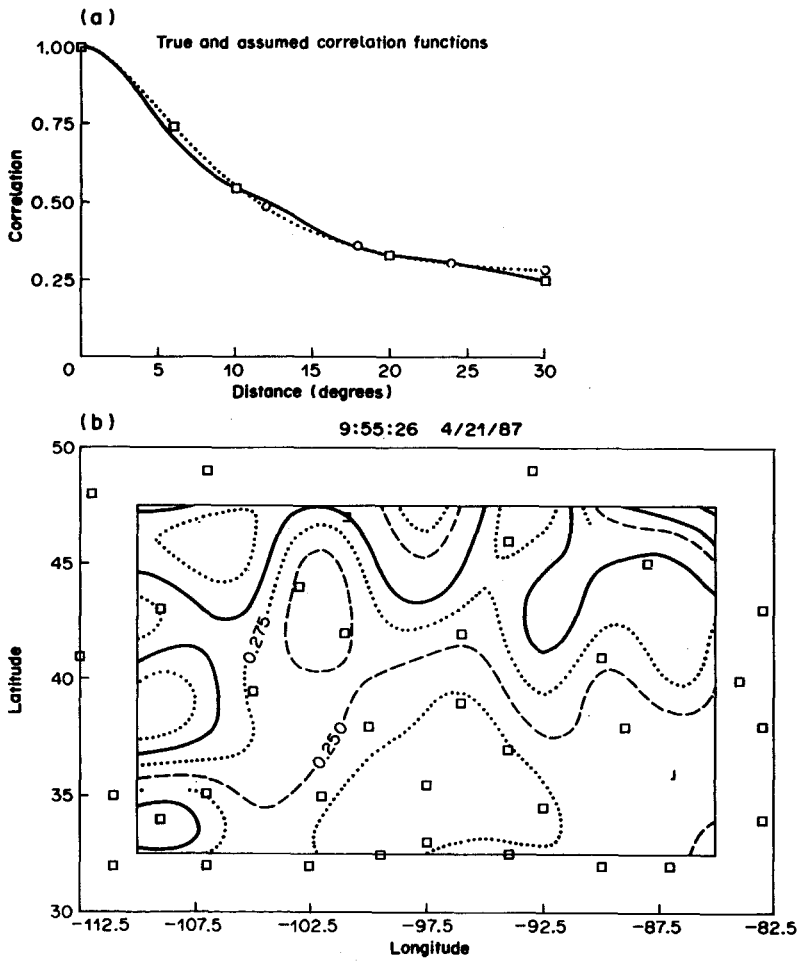

Fig. 3. (a) Six-term Bessel correlation function [true (□)] and second-order autoregressive plus constant correlation function [assumed (○: AR2 0.2055; 0.2722)]. (b), (c) and (d), as in corresponding parts of Fig. 2 for the correlation functions shown in part (a). AS = 2:0 0.206 0.272; OBS/FG = 0.333.

in Table 1 show very close agreement with the optimum possible for all three of the grid/observation sets.

(*v*) *Damped cosine.* The damped cosine function was suggested by Thiébaux [19] and Seaman and Hutchinson [34]. The formula is

$$[A + (1 - A)\cos(as)]/[(1 + (bs)^2]^c. \tag{9}$$

It is unknown whether the function is positive definite as an isotropic function in two dimensions, but the evidence in Section 2.4 (for $c = 0.5$), while inconclusive, seems to indicate it is not. In practice, of course, the function may be positive definite when the observation points are restricted to certain regions. The data was fitted with function (9), under the restriction $c = 0.5$. The intercept was 0.8565, which resulted in a very slight raising of the curve relative to the model correlation function. The resulting fit is excellent for small distances and very good over the entire range. Table 1 shows that this function gives the best results of all the functions tested.

(*vi*) *Variations.* The expected error computations for a number of variations of the above functions were also performed. The principal variation was to fit the data only over the first half of the interval, (0°, 15°). The effect of this was to generally (though not always) increase the e.r.m.s. errors over MUS and EC, while not affecting the results over MA. In the damped cosine, the exponent c was chosen by least squares, along with the other parameters, and resulted in a slightly better fit to the correlation function, especially at larger distances. However, the coefficient A was slightly greater than unity. Whatever the positive definiteness properties of the function, having $A > 1$ will certainly make it nonpositive definite. Although no graphical results are shown, the coefficients and e.r.m.s. errors are given in Table 1 for the additional assumed correlation functions.

3.5. Sensitivity of the SOAR model to parameter misspecification

In order to determine more completely the characteristics of the SOAR model, some additional calculations were made to determine the effect of misspecification of the parameters in the correlation function or the ratio of the standard deviations of the observed and background error. The results can be summed up rather quickly: the scheme is mostly insensitive to such variations. Figure 4 shows a family of four correlation functions, No. 4 being the SOAR plus constant discussed in the Section 3.4, with the others having smaller correlations at a given distance. Figure 5 shows the e.r.m.s. error for each of the four as the "assumed" correlation, when the "true" correlation function is No. 4. With the exception of the sparse MA grid, the expected errors are relatively stable under significant perturbations. Figure 6 shows the sensitivity to the assumed error ratio, and once again, it is observed that the expected errors are quite stable.

4. CONCLUSIONS

The principal conclusion to be drawn is that the correlation family used in practical analysis should embody a sufficient number of parameters to fit the forecast error statistics reasonably well. Further, it is most important that the data be fitted accurately for small distances. In order to

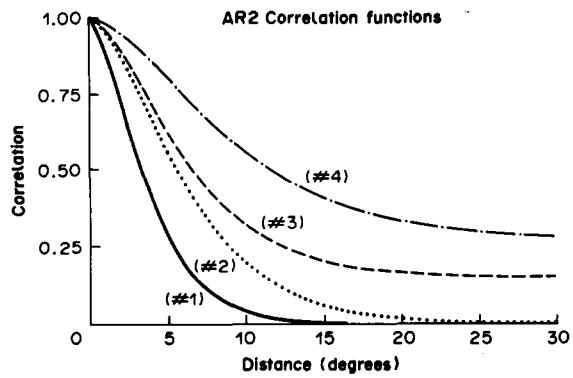

Fig. 4. Four second-order autoregressive correlation functions, as in equation (7a), with (b, A) values: 1—(0.5, 0.0); 2—(0.3, 0.0); 3—(0.3 = 0.15); 4—(0.2055, 0.2722).

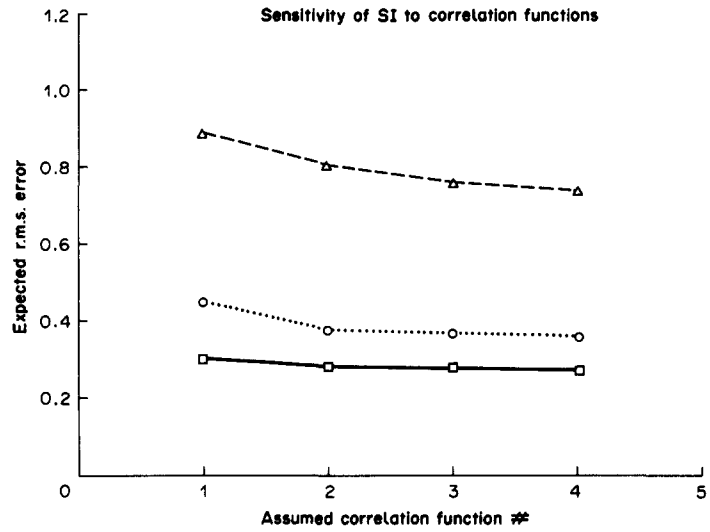

Fig. 5. e.r.m.s. errors when the "true" correlation is function No. 4 with various assumed correlation functions for each of the three grid/observation sets. (□) MUS; (○) EC; (△) MA.

ensure a better fit for small distances, it may be worthwhile to enforce the intercept of the correlation for the background plus observation errors *if* the ratio of standard deviations of the two errors is known *accurately*. The effect of scaling to obtain the correlation function, and the apparent shift up or down can possibly be compensated for by artificially varying the ratio of background/observation errors, as well, although it seems more desirable to enforce this ratio in the correlation function fitting process.

As noted above, clearly the most important region for the fit to the correlation function to be accurate is for small distances. Over the sparsely observed region, MA, and to a lesser extent over the EC region, the overall e.r.m.s. errors were only slightly affected by the assumed correlation function. In the case of the MA region it is noted that the error contours are relatively unaffected, except near the observations. Since the errors in the remote part of the region dominate the overall error, the choice of assumed correlation function has relatively small influence. On the other hand, over the densely observed region, an accurate fit at small distances was most important. The NSE correlation function, while not performing well, illustrates the above nicely. For the first NSE entry

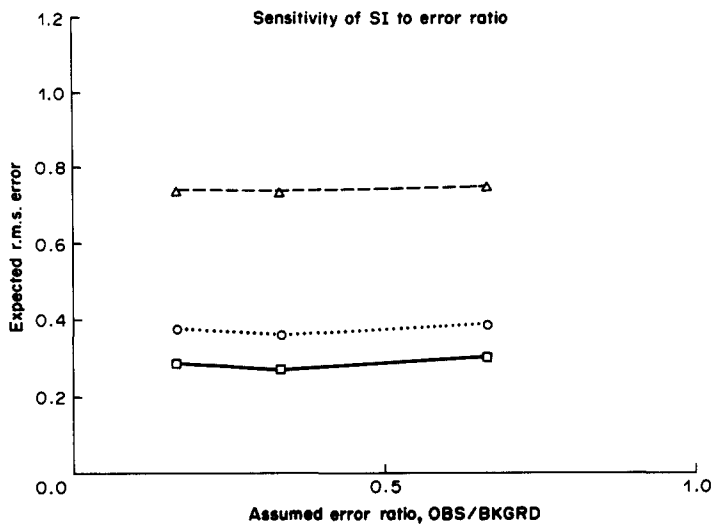

Fig. 6. e.r.m.s. errors when the assumed ratio of the standard deviations of the observed to background error is varied. Actual error ratio is 1/3. (□) MUS; (○) EC; (△) MA.

in Table 1, even though the fit is poor for distances $>6°$, the e.r.m.s. errors over MUS and EC are smaller than the best fit (last NSE entry) due to the more accurate fit for small distances by the former function. Of course the e.r.m.s. errors over MA are poorer for the first case due to the very bad fit at large distances.

There appear to be several good candidates for use as two-dimensional isotropic correlation functions, including the SOAR, the TOAR, and the damped cosine, given by equations (7), (8) and (9), respectively. While the fitting power for the latter two are greater (there are a greater number of parameters for those two), the choice of SOAR seems reasonable and adequate for a number of reasons:

(1) The SOAR (with the additive constant) embodies a sufficient number of parameters to allow oscillation and decay with distance.
(2) The SOAR has some credibility as the spatial correlation function of an innovation process. However, results are for one dimension rather than two, except for the results cited previously in Balgovind *et al.* [33].
(3) The SOAR was demonstrated here to be positive definite as an isotropic function in two dimensions, under a mild restriction on the parameters.
(4) While the TOAR is also the spatial correlation function (again in one dimension) for an innovation process, based on this limited study it does not appear to be significantly better than the SOAR.
(5) The positive definiteness properties of the TOAR are not known, although it is certainly positive definite as an isotropic function in two dimensions under some restrictions on the parameters.
(6) Although the fitting ability of the damped cosine seems to be at least as good as the TOAR, and it is positive definite in one dimension, evidence indicates it may not be positive definite as an isotropic function in two dimensions, regardless of parameter restrictions. The availability of other acceptable alternatives seems to make it prudent to preclude the use of the damped cosine in practical situations.

It is pointed out that all of the functions except the four-term Bessel function and the NSE perform very well. Table 1 shows, for example, that the SOAR is only a little more than 1% of the standard deviation of the background error poorer than optimal over MUS and EC, and $<0.1\%$ poorer over MA.

Finally, it is noted that within the SOAR family, SI is quite insensitive to misspecification of the correlation parameters, even to an extent such that the correspondence would appear to be much less between two members of the family than between it and a fit by the NSE. Thus it could as important to choose the correct family of correlation functions as well as to model properly within that family. In addition, misspecification of the ratio of standard deviations of the background and observation errors has a rather small effect on the skill of the method.

This work has focused only on the univariate problem, whereas in practice such schemes are applied to the multivariate one. Further work is necessary to determine whether the nice results obtained here carry over to the multivariate case. A further investigation of the effect of wind observations on the analysis of pressure height and wind fields is anticipated.

Acknowledgement—R. Franke was supported by the Office of Naval Research under Program Element 61153N, Project No. BR033-02-WH.

REFERENCES

1. V. Bjerknes, Das problem der wettervorhersage betrachtet von standpunkt der mechanik und der physik. *Met. Z.* **21**, 1–6 (1904).
2. L. F. Richardson, *Weather Prediction by Numerical Processes*. Dover, New York (1922). Copyright 1965.
3. J. G. Charney, R. Fjortoft and J. von Neuman, Numerical integration of the barotropic vorticity equation. *Tellus* **2**, 237–254 (1950).
4. H. Panofsky, Objective weather map analyses. *J. Met.* **6**, 386–392 (1949).
5. P. Bergthorsson and B. R. Döös, Numerical weather-map analysis. *Tellus* **7**, 329–340 (1955).
6. G. Cressman, An operational objective analysis system. *Mon. Weath. Rev.* **87**, 367–374 (1959).
7. L. S. Gandin, Objective analysis of meteorological fields (1963). Translated from Russian by the Israel Program for Scientific Translations, NTIS TT65-50007 (1965).

8. I. D. Rutherford, Data assimilation by statistical interpolation of forecast error fields. *J. atmos. Sci.* **29,** 809–815 (1972).
9. T. W. Schlatter, Some experiments with a multivariate statistical objective analysis scheme. *Mon. Weath. Rev.* **103,** 246–257 (1975).
10. T. W. Schlatter, G. W. Branstator and L. G. Thiel, Testing a global multivariate statistical objective analysis scheme with observed data. *Mon. Weath. Rev.* **104,** 765–783 (1976).
11. K. H. Bergman, Multivariate analysis of temperatures and winds using optimum interpolation. *Mon. Weath. Rev.* **107,** 1423–1444 (1979).
12. A. C. Lorenc, A global three dimensional multivariate statistical interpolation scheme. *Mon. Weath. Rev.* **109,** 701–702 (1981).
13. H. J. Thiébaux, On approximations to geopotential and wind–field correlation structures. *Tellus* **37A,** 126–131 (1985).
14. H. J. Thiébaux and M. A. Pedder, *Spatial Objective Analysis.* Academic Press, London (1986).
15. S. Bochner, *Lectures on Fourier Integrals* (Translated from the original by M. Tenenbaum and H. Pollard). Princeton Univ. Press, Princeton, N.J. (1959).
16. J. Stewart, Positive definite functions and generalizations, an historical survey. *Rocky Mount. J. Math.* **6,** 409–434 (1976).
17. C. E. Buell, Correlation functions for wind and geopotential on isobaric surfaces. *J. appl. Met.* **11,** 51–59 (1972).
18. P. R. Julian and H. J. Thiébaux, On some properties of correlation functions used in optimum interpolation schemes. *Mon. Weath. Rev.* **103,** 605–616 (1975).
19. H. J. Thiébaux, Anisotropic correlation functions for objective analysis. *Mon. Weath. Rev.* **104,** 994–1002 (1976).
20. H. J. Thiébaux, Extending estimation accuracy with anisotropic interpretation. *Mon. Weath. Rev.* **105,** 691–699 (1977).
21. H. J. Thiébaux, H. L. Mitchell and D. W. Shantz, Horizontal structure of hemispheric forecast error correlations. In *Preprints 7th Conf. on Numerical Weather Prediction,* Montreal, P.Q., pp. 17–26 (1985).
22. R. S. Seaman, A systematic description of the spatial variability of geopotential and temperature in the Australian region. *Aust. met. Mag.* **30,** 133–141 (1983).
23. C. E. Buell and R. S. Seaman, The "scissors" effect: anisotropic and ageostrophic influences on wind correlation coefficients. *Aust. met. Mag.* **31,** 77–83 (1983).
24. B. D. Ripley, *Spatial Statistics.* Wiley, New York (1981).
25. B. Matèrn, *Spatial Variation,* Meddelanden fran Statens Skogsforskningsinstitut 49, 5, pp. 1–144 (1960).
26. G. Matheron, The intrinsic random functions and their applications. *Adv. appl. Probabil.* **5,** 439–468 (1973).
27. H. Hochstadt, *Integral Equations.* Wiley, New York (1973).
28. R. Franke, Covariance functions for statistical interpolation. Technical Report NPS-53-86-007, Naval Postgraduate School, Monterey, Calif. (1986).
29. P. Lonnberg, Structure functions and their implications for higher resolution analysis. In *Proc. Wkshp on Current Problems in Data Assimilation,* ECMWF, pp. 142–178 (1982).
30. G. Wahba and J. Wendelberger, Some new mathematical-methods for variational objective analysis using splines and cross validation. *Mon. Weath. Rev.* **108,** 1122–1143 (1980).
31. M. Ghil, S. Cohn, J. Tavantzis, K. Bube and E. Isaacson, Applications of estimation theory to numerical weather prediction. In *Dynamic Meteorology: Data Assimilation Methods* (Edited by L. Bengtsson, M. Ghil and E. Kallen), pp. 139–224. Springer, New York (1981).
32. M. I. Yudin, Some regularities in the structure of the geopotential field. *Trudy GGO* **121,** 3–18 (1961).
33. R. Balgovind, A. Dalcher, M. Ghil and E. Kalnay, A stochastic–dynamic model for the spectral structure of forecast error statistics. *Mon. Weath. Rev.* **111,** 701–722 (1983).
34. R. S. Seaman and M. F. Hutchinson, Comparative real data test of some objective analysis methods by withholding observations. *Aust. met. Mag.* **33,** 37–46 (1985).